Lexical bundles and discourse signaling in academic lecturers.

Nesi, H. and Basturkmen, H.

Original citation & hyperlink:
Nesi, H. and Basturkmen, H. (2006) Lexical bundles and discourse signaling in academic lecturers.. International Journal of Corpus Linguistics, volume 11 (3): 283-304.
http://www.benjamins.com/cgi-bin/t_articles.cgi?bookid=IJCL%2011%3A3&artid=659063338

Nesi, H. and Basturkmen, H. (2006) Lexical bundles and discourse signaling in academic lecturers. *International Journal of Corpus Linguistics*, volume 11 (3): 283-304

Abstract
This paper discusses some approaches to the categorisation of cohesive devices with reference to spoken academic discourse, multi-word units, and strings of frequently co-occurring words (lexical bundles). It goes on to investigate the cohesive role of lexical bundles in a corpus of 160 university lectures (120 from the BASE corpus and 40 from MICASE). Like the bundles from the T2K SWAL teaching subcorpus, investigated by Biber, Conrad and Cortes (2004), the bundles in the lecture corpus included both 'oral' and 'literate' elements. The majority of frequently occurring bundles were found to be used to signal discourse relations, although their cohesive function was not necessarily obvious when listed out of context.

1. Introduction

1.1 Cohesion and multiword units

Halliday and Hasan's seminal work on cohesion in English (1976) identified four broad categories of textual relation: 'addition', 'comparison', 'time sequencing' and 'result'. Subsequent researchers have developed this system, altering the terminology and categories to suit their individual purposes, but although there are many terminological differences between one system and another, the categories of additive, adversative, sequential and causal relations are common to most classifications (Louwerse & Mitchell 2003). The *Longman Grammar of Spoken and Written English* (Biber et al. 1999) distinguishes between 'coordinators', 'subordinators' and 'linking adverbials' as grammatical devices which explicitly signal the connections between passages of text. While coordinators and subordinators signal meaning relations within sentences, linking adverbials signal intrasentential relations, and can be compared to 'cohesive conjunctions', as described by Halliday and Hasan (1976). Biber et al. (1999) divide linking adverbials into six semantic groups: enumeration and addition (to list and add items to the ongoing discourse), summation (to conclude preceding discourse), apposition (to indicate equivalence or inclusion), result/inference (to signal consequence), contrast/concession (to indicate dissimilar comparison) and transition (to signal asides and lack of continuity with preceding discourse). Biber et al. (1999) report differing patterns of cohesion in conversation, academic prose, fiction and news reports, finding that linking adverbials occur more often in conversation and academic prose than in fiction and news. Adverbials signalling enumerative/summative and appositional relations are most common in academic prose.

Many of the items that have been identified as serving a signalling function in discourse are multi-word units rather than single words. Of course, some of the adverbs that typify cohesive conjunctions (Halliday & Hasan 1976) originally presented as two- or three-word units (*furthermore, however, nevertheless* etc), but came to be regarded as single words as they acquired a conventionalised pragmatic meaning. (Other such compounds seem to be in the process of transition — *inasmuch as, insofar as, instead of* etc.) In Halliday and Hasan's scheme, conjunction is realised by either single word adverbs or multi-word prepositional phrases (such as *in addition, as a result of that, in spite of that*), but most recent research admits a broader

range of syntactic structures into the cohesive system. Biber et al. (1999), for example, include within their category of 'linking adverbials' *which is to say* (as an adverbial of apposition), and *to conclude* (as an adverbial marking summation). Hyland (2004) categorises *while it is true* as a conjunction indicating concession.

Multi-word discourse devices are typically identified by combining text analysis with intuition about the way words are used. DeCarrico and Nattinger (1988) and Nattinger and DeCarrico (1992) examined the collocations within a span of five lexical items surrounding pre-selected node words in a corpus of lectures. They counted as lexical phrases those groups of words which appeared to be prefabricated, rather than generated by syntactic competence, and which served a pragmatic function in the discourse. A variety of multiword discourse devices were identified by this method, but the identification process required the researchers first of all to notice the existence of a prefabricated chunk and define its boundaries. For this reason Nattinger and DeCarrico's lexical phrases tend to be perceptually salient, and structurally complete.

1.2 The identification and description of lexical bundles

An alternative method of identifying multi-word units is empirical rather than intuitive. Strings of frequently co-occurring words can be identified within a given corpus regardless of syntactic boundaries or their salience as meaningful units independent of context. Strings of any given length can be identified, and their meaning and function can be considered subsequently. Groups identified by this means are referred to by a variety of different terms, as Stubbs (2002:230) points out: they are known as 'clusters' (Scott 1997:41), 'recurrent word-combinations' (Altenberg 1998:101), 'statistical phrases' (Strzalkowski 1998:xiv), 'lexical bundles' (Biber et al. 1999:993) and 'n-grams' (Banerjee & Pedersen 2003).

Lexical bundles are discussed at length in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999, Section 13.2). In this work a bundle is defined as 'a recurring sequence of three or more words' (1999:90), although most of Biber et al.'s analysis concerns longer sequences, particularly the fourword bundle. Biber et al. found almost ten times as many three-word bundles as four-word bundles in their corpus of conversation and academic prose, and almost ten times as many four-word bundles as five-word bundles. Four-word bundles included extended versions of the most common three-word bundles, and, as they were plentiful, they restricted their analysis to those occurring ten or more times per million words, in five or more different texts, in order 'to exclude individual speaker/writer idiosyncracies'(Biber et al. 1999:993).

One disadvantage of this method is that it does not permit the identification of discontinuous frames (for example *not only… but also…*). The pre-specification of string length does not preclude the identification of longer strings, however, as there is usually some overlap between the most frequently occurring strings of any given length. As the *Longman Grammar of Spoken and Written English* points out, 'longer lexical bundles are usually formed through an extension or combination of one or more shorter bundles' (Biber et al. 1999:993). By way of illustration, Biber et al. show the development of a string from three to six words:

do you want; you want me; want me to; me to do →
do you want me; you want me to; want me to do →
do you want me to; you want me to do →
do you want me to do.

Later studies have continued the practice of focussing on four-word bundles, often setting more rigorous cut-off points. Cortes (2004), for example, only considers bundles occurring 20 or more times per million words, and Biber, Conrad and Cortes (2004) only examine those occurring 40 times per million. Lexical bundles that occur with very high frequency across a range of texts are likely to be stored in memory as unanalysed chunks, a particularly interesting consideration in view of the fact that bundles tend to bridge syntactic boundaries and do not generally have idiomatic meaning, and are therefore not very salient, either to the listener/reader or to the language researcher. As Biber, Conrad and Cortes (2004:377) point out:

> for the most part linguists have not noticed these high frequency multi-word sequences, probably because most previous research has focussed on grammatical phrases and clauses, disregarding the possibility of lexical units that cut across grammatical structures.

Biber, Conrad and Cortes (2004) compared lexical bundles in academic prose, conversation and teaching sessions, using text samples taken from the *Longman Spoken and Written English Corpus* and the *TOEFL 2000 Spoken and Written Academic Language Corpus* (T2K SWAL). Previous work on lexical bundles had concentrated on written text and conversation, but Biber, Conrad and Cortes (2004:382) found that classroom discourse made far more frequent use of bundles, and contained the greatest variety of different bundle types. They concluded that 'the extremely high density of lexical bundles in classroom teaching exists because this register relies heavily on both 'oral' and 'literate' bundles. In their data 'oral' bundles, typical of conversation, are characterised by declarative and interrogative clause fragments, while 'literate' bundles, typical of academic prose, contain noun phrases and prepositional phrases. Their teaching subcorpus bundles contained more nouns than the bundles in the conversation subcorpus, but more verbs than the bundles in the subcorpora of textbooks and academic prose.

1.3 The cohesive role of lexical bundles

Biber, Conrad and Cortes (2004) do not explicitly discuss cohesion with reference to lexical bundles, and do not provide many examples of bundles which perform a cohesive role. Nevertheless some of the functions of lexical bundles they describe seem comparable to those assigned to conjunctive relations in the model of cohesion proposed by Halliday and Hasan (1976), and to functions identified in other accounts of metadiscourse in academic text (for example Hyland 2004). Three primary functions are identified: I. stance (expressing attitudes or assessments of certainty), II. discourse organization (reflecting relationships between prior and coming discourse), and III. reference (referring to physical or abstract entities, or to the textual context). The bundles expressing function II and (to a lesser extent) function III appear to have the greatest potential to perform a cohesive role. Bundles expressing function II are subdivided into two further categories: A, topic introduction/focus, and B, topic elaboration/clarification. Category IIA includes such bundles as *if you look at*; *going*

to talk about; *what I want to* etc. These can be equated with 'frame markers' in Hyland's model of interactive metadiscourse in academic texts (Hyland 2004), and mark transition points in text. It is possible that they play a cohesive role as continuative items, fulfilling similar functions to *now* or *well* in the semantic system described by Halliday and Hasan (1976). Category IIB includes such bundles as *has to do with*; *you know I mean*; *on the other hand*; *as well as the* etc. These can be equated with 'transitions', the items in Hyland's model which are used to signal semantic relations between main clauses (Hyland 2004). These items likewise express some of the conjunctive relations identified by Halliday and Hasan (1976). Bundles expressing function III are subdivided into four further categories: A, identification/ focus, B, imprecision, C, specification of attributes, and D, time/place/text reference. Most of the bundles assigned to categories IIIB (eg *and stuff like that*) and IIIC (eg *the size of the, the nature of the*) do not fit into Hyland's model of metadiscourse in academic texts (Hyland 2004) and do not appear to perform a cohesive function (with the exception of *as a result of*, which apparently marks a causal conjunctive relation). Many of the bundles in category IIIA, however, seem to signal exemplification (*that's one of the, one of the things, of the things that* etc) and might thus perform a cohesive role similar to that of 'additive apposition' in the system proposed by Halliday and Hasan (1976). Category IIID includes some items that might function as 'frame markers' in Hyland's model (*at the end of the, the beginning of the* etc) (Hyland 2004), and might perform the same role as temporal conjunctions (Halliday & Hasan 1976).

1.4 Hypotheses and research questions

Biber, Conrad and Cortes (2004) claim knowledge of only one study prior to theirs which examines the use of multi-word units in university lectures. This is the work of DeCarrico and Nattinger (1988) (also reported in Nattinger & DeCarrico 1992), which, as discussed earlier, identified units primarily on the basis of perceptual salience, and therefore tended to ignore those frequently occurring clusters that are syntactically incomplete, and composed of very common words which do not take on any special idiomatic meanings when they co-occur.

Biber, Conrad and Cortes (2004) examined lexical bundles that had been identified automatically, in the manner described in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). However most of the 'lectures' in their teaching subcorpus did not represent monologic, lecture-style discourse, because in T2K SWAL about three quarters of the recorded teaching sessions are more like lessons than lectures, with many short turns of less than 100 words (Biber, Conrad, Reppen, Byrd, Helt, Clark, Cortes, Csomay & Urzua 2004). This kind of classroom discourse is characterised by 'interactions among participants, and a focus on the speakers' personal concerns' (Biber, Conrad & Cortes 2004:378), leading to more 'oral' lexical bundles. Lectures, on the other hand, are less interactive and more pre-planned, and in this respect may be closer to academic writing.

The circumstances that Biber, Conrad and Cortes (2004) consider responsible for the unusually high number of lexical bundles in their teaching subcorpus should also apply to monologic lectures, however. Like participants in conversation, both lecturers and teachers face real-time production constraints which encourage the use of prefabricated chunks. Moreover in both teaching contexts the information content is high and the pedagogic function will require that connections between propositions are

made clear, although the fact that pre-planned monologic lectures are not co-constructed, and lack normal opportunities for negotiation of meaning, suggests that they may have a greater need for discourse structuring devices. Because of the differences between our lecture data and classroom discourse, and because of our particular focus on cohesion, we expect in this study to shed light on some devices that have not been discussed in previous accounts of the cohesive system.

As an exploration of the cohesive role of lexical bundles in a corpus of lectures, this study addresses two research questions:

1. How do the lexical bundles used in lectures compare to those in other registers, and particularly to those used in classroom teaching?
2. How do the lexical bundles used in lectures help to create cohesion?

2. Procedure

Our corpus was made up of 160 monologic lectures from the *British Academic Spoken English* (BASE) corpus1 and the *Michigan Corpus of Academic Spoken English* (MICASE).2 To create this corpus thirty lectures from BASE and ten from MICASE were selected from each of four broad disciplinary groupings: Arts and Humanities (325,873 words), Social Sciences (360,037 words), Life Sciences (316,762 words) and Physical Sciences (268,126 words). The total corpus size was 1,270,798 words, similar to that of the classroom teaching subcorpus used by Biber, Conrad and Cortes (2004) (1,248,800 words).

The BASE component was considerably larger than the MICASE component (882,980 words as opposed to 387,818 words) because although the two complete corpora are roughly equivalent in size, MICASE covers a wider range of speech event types and contains fewer examples of lectures dominated by a single speaker.

Lexical bundles in this corpus were identified using WordSmith Tools (Scott 1997). WordList cluster processing was activated by choosing the option *Settings/Min. & Max. Frequencies* from the main WordList menu. A cluster size of four words was selected, with a minimum frequency of ten.

3. Results

3.1 The frequency of lexical bundles in lectures

Most of the commonest bundles occurred with very similar frequency in both the British and American lectures, although *in the United States* (relatively rare in BASE) was the 15th most frequent bundle in the MICASE component, and *in the UK* (relatively rare in MICASE) was the 10th most frequent in BASE. Slight differences in the way the two corpora were transcribed may have affected findings to a small extent: MICASE makes greater use of contracted forms such as *gonna* and *wanna* (the bundle *we're gonna talk about* occurred 24 times in the MICASE sample) while the BASE transcribers were more likely to use *going to* and *want to*. Fillers and hesitation markers are marked as # in the BASE corpus and represented lexically in MICASE as *hm, huh, mm, mhm, uh, um* etc. However none of the most frequent bundles in either corpus contained markers of this kind, a further indication of their prefabricated nature.

There were 33,761 instances of four word bundles occurring at least ten times in the corpus as a whole. One thousand two hundred and sixty-six different four word bundles occurred ten times or more — about 996 per million words. Of these, 34 occurred at least 60 times — roughly 27 per million words. This finding is almost identical to that of Biber, Conrad and Cortes (2004); 28 four-word bundles occurred at least 60 times per million words in their teaching subcorpus, as opposed to 44 in conversation. Like classroom teaching, lectures use 'a large set of different lexical bundles, while conversation relies on the extremely frequent use of a smaller set of bundles' (Biber, Conrad & Cortes 2004:379).

3.2 The characteristics of lexical bundles in lectures

Table 1 compares the word classes of lecture bundle endings with those occurring in conversation and academic writing, as described in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999:997 Table 13.3).

Word class	Example	Lectures	Conversation	Academic prose
Verb	*I want to know*	17%	40%	5%
Pronoun	*well that's what I*	11%	15%	–
other function word	*I went to the/in the case of*	50%	40%	85%
Noun	*at the same time*	17%	–	10%

Table 1: The grammatical category of words ending bundles in lectures, conversation and academic prose (approximate proportional distribution)

Like the bundles in the teaching subcorpus, the lecture bundles included both 'oral' and 'literate' elements. Table 2 reveals in more detail the mixed 'oral' and 'literate' nature of the lecture bundles. Their proportional distribution is compared with that of bundles in conversation and academic writing, as described in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999:996 Table 13.1). Biber, Conrad and Cortes (2004) found that whereas conversation made greater use of VP-based and dependent clause bundles, and textbooks made greater use of NP/PP-based bundles, classroom teaching made roughly equal use of all three kinds of bundles. The first four structures in Table 2 are commoner in conversation than in academic writing, and the second four structures are commoner in academic writing than in conversation. They all occur with some frequency in the lecture corpus, although question fragments are not well represented. In our corpus questions sometimes serve as a means of structuring discourse and managing topic change, but there is very little use of the questioning strategies typical of classroom interaction (as schematised by Sinclair & Coulthard (1975) in their Initiation Response Feedback text pattern). In contrast Biber, Conrad and Cortes (2004:382) report that classroom teaching 'makes dense use of lexical bundles that represent declarative and interrogative clause fragments'.

Structures within the 'other expressions' category in Table 2 included four word numerical expressions, long nominal compounds and adverbial clause fragments such as *if*-clauses.

Structure	Example	Lectures	Conver-sation	Academic prose
Personal pronoun + lexical verb phrase (+ complement clause)	*I don't know what*	15%	44%	–
Pronoun/NP (+auxiliary) + copula *be* (+)	*it was in the*	10%	8%	2%
(auxiliary +) active verb (+)	*have a look at*	14%	13%	–
Yes-no and *wh*-question fragment	*can I have a*	2%	12%	–
(verb +) *wh*-clause fragment	*know what I mean*	4%	4%	–
Noun phrase with post-modifier fragment	*the nature of the*	19%	4%	30%
Preposition + noun phrase fragment	*as a result of*	9%	3%	33%
Anticipatory *it* + vp/adjective P (+complement clause)	*it is possible to*	1%	–	9%
Passive verb + PP fragment	*is based on the*	–	–	6%
(verb +) *that*-clause fragment	*should be noted that*	3%	1%	5%
(verb/adjective +) *to*-clause fragment	*are likely to be*	3%	5%	9%
Other expressions		17%	6%	6%

Table 2: Structural patterns in lectures, conversation and academic prose (approximate proportional distribution)

We examined in particular detail those four word lexical clusters that occurred at least ten times within each disciplinary grouping, and over 50 times in the corpus as a whole. By selecting according to range of occurrence as well as frequency we eliminated from our study numerical sequences and technical terms (e.g. *the D N A* and *R N A Polymerase*, both of which occurred 50 times in the corpus as a whole, but only in one disciplinary grouping). By this means we also eliminated any clusters that were idiosyncratic to a single speaker.

Table 3 shows the 20 most frequent lexical bundles in the corpus, all of which occurred at least ten times in each broad disciplinary grouping. Four pairs of bundles in this list can be combined to form some of the most frequently occurring five word bundles:

1. *the end of the; at the end of → at the end of the* (occurring 84 times)
2. *if you look at; you look at the → if you look at the* (occurring 56 times)
3. *one of the things; of the things that → one of the things that* (occurring 55 times)
4. *and you can see; you can see that → and you can see that* (occurring 27 times).

Seventeen of these 20 most frequent bundles are listed by Biber, Conrad and Cortes (2004:384–8) as common bundles in classroom teaching. Three did not occur frequently in any of their subcorpora (*you look at the*; *and you can see; you can see that*).

		Total	MICASE	BASE	Life	Arts	Social	Physical
1.	THE END OF THE	171	44	127	29	76	36	30
2.	AT THE END OF	152	37	115	35	57	31	29
3.	IF YOU LOOK AT	132	38	94	30	32	43	27
4.	TO BE ABLE TO	128	43	85	43	19	36	30
5.	AT THE SAME TIME	118	36	82	24	42	35	17
6.	IF YOU WANT TO	113	13	100	31	15	32	35
7.	ONE OF THE THINGS	110	43	67	34	29	33	14
8.	IN TERMS OF THE	109	29	80	23	22	45	19
9.	IS GOING TO BE	104	14	90	18	11	17	58
10.	AND YOU CAN SEE	91	22	69	28	22	18	23
11.	A LITTLE BIT OF	83	32	51	37	10	19	17
12.	YOU CAN SEE THAT	82	13	69	31	18	16	17
13.	OF THE THINGS THAT	76	31	45	31	12	21	13
14.	THE REST OF THE	75	34	41	27	23	14	11
15.	AND THIS IS THE	74	18	56	30	16	12	16
16.	I WANT YOU TO	72	30	42	16	17	21	18
17.	YOU LOOK AT THE	69	16	53	21	15	20	13
18.	I'M NOT GOING TO	68	6	62	20	11	17	20
19.	TO DO WITH THE	65	18	47	13	21	21	10
20.	A LOT OF THE	63	14	49	15	19	19	10

Table 3: Frequent lexical bundles

3.3 The cohesive role of lexical bundles in lectures

The objective of the next stage of the research was to investigate whether the bundles played a discourse signalling role. The concordance function from the software package MP 2.2 (Barlow 2002) was used to retrieve and display the concordances surrounding the occurrences of these bundles. We used a method similar to that used by Flowerdew (2003) in an investigation of signalling nouns in biology textbooks and lectures. A qualitative analysis of the concordance lines displayed on the computer screen was conducted. We examined the concordances to identify whether in some instances the bundles appeared to function to link parts of the discourse, and when this appeared to be the case the software was used to retrieve a longer section of co-text. In general this involved tracking the text back to the point where the topic of the text segment in which the lexical bundles appeared was introduced into the discourse, and tracking the text forward to the point where the second idea unit ended. This long section was then examined with the aim of identifying the type of cohesive relation involved. We made use of the semantic categories of linking adverbials listed in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999): enumeration/addition, summation/conclusion, apposition, result/inference, contrast/concession and transition.

To help determine the type of relationship involved we tried, on an intuitive basis, to replace the bundles with tokens widely recognised as signalling different types of relationship between parts of text, such as *in addition*, *so*, *in other words* and *by the*

way. No attempt was made to identify all instances in the data when lexical bundles appeared to signal discourse relationships, or quantify the extent to which bundles perform this role. Our aim was to establish whether bundles had the potential to contribute to the cohesive system in lectures, and to draw attention to some of the ways in which this might be achieved.

With the exception of *is going to be* and *the rest of the*, all the frequent lexical bundles listed in Table 3 were observed to function at times to signal discourse relations. In other contexts, all the bundles also served other functions, often as directives or as expressions of stance. The bundles *if you look at/you look at the → if you look at the*, for example, functioned variously as directives, 'topic introduction/ focus' discourse organizers, and referential expressions (signalling exemplification). Biber, Conrad and Cortes (2004:383) point out that bundles can have multiple functions even within a single occurrence.

In this article there is only space for a few examples to show the main cohesive relationships signalled by bundles in the corpus. An examination of the bundles in their surrounding text revealed that their discourse signalling functions were of two kinds. They appeared firstly to signal how one idea or piece of information was related to another idea or piece of information in the lecture. In the discussion below such signals will be referred to as 'referential expressions' (borrowing the term used for category III markers in Biber, Conrad & Cortes 2004). They appeared secondly to signal the relationship between topics and activities in the lecture, roughly equivalent in function to 'frame markers' (Hyland 2004:138), which are references to text boundaries, and devices used to sequence and label stages in the text and announce changes in topic and discourse goals. In the following discussion, markers with this function will be referred to as 'discourse organizers' (borrowing the term used for category II markers in Biber, Conrad & Cortes 2004).

3.3.1 *Referential expressions*

Bundles sometimes seemed to indicate logical relationships such as apposition, contrast/concession and result/inference (all semantic categories of linking adverbials in the *Longman Grammar of Spoken and Written English*, Biber et al. 1999).

The relationship of apposition is described in the *Longman Grammar of Spoken and Written English* as "showing that the second unit of text is to be treated either as equivalent to or included in the preceding unit… an appositive linking adverbial can be used to show that the second unit is to be taken as a restatement of the first, reformulating the information it expresses in some way or stating it in more explicit terms" (Biber et al. 1999:876). In our corpus we found appositive relations signalled by the bundles *and you can see*/*you can see that*; *if you look at/you look at the; one of the things/of the things that; a lot of the;* and *and this is the.* Examples 1 to 3 illustrate propositions related by apposition. After each example, one of the linking adverbials listed in the *Longman Grammar of Spoken and Written English* under apposition is provided. It is suggested that these adverbials could replace the lexical bundles in the extracts with small changes to the sentence structure but almost no change in meaning.

Example 1

He takes the name victorious general that's what imperator means so you can see now when you meet him on the street you could say oh how are you you are the victorious general son of a god Caesar how very nice for you *and you can see* how Roman names can be used to carry a message (in other words)

Example 2
They got a B on a paper they thought they were going to get an A on and then miss something like a thyroid diagnosis so it's really important to get a grip on the physical *one of the things that* is very easy to forget to ask about especially when you are confronted with a person who is clean cut looks good very pulled together very articulate is to ask about drug and alcohol abuse (for example)

Example 3
There is one more polarity which you need to register it's crucial because it lies at the very heart of Huck's moral dilemma *and this is the* opposition between conscience and heart (namely).

In our corpus it was not uncommon for lecturers to use frequently occurring lexical bundles alongside other items conventionally recognised as cohesive conjunctions. The following two examples show this; apposition is signalled by both a lexical bundle (*if you look at*) and a linking adverbial (*for example* in Example 4, and *for instance* in Example 5).

Example 4
well there is lots of apparent evidence which has been interpreted as evidence of social learning *for example if you look at* different colonies or troops of chimps the same species but living in different parts of Africa you find that different ch- troops show different behaviours

Example 5
once you've got them and you've got some numbers on the diagram you can decide which are the sort of dominant issues and if something isn't going to be important in the overall heat transfer *for instance* in this problem *if you look at* the thermal resistance of the glass it's very very small so in terms of changing the heat transfer it really wouldn't matter if you made the glass a bit thicker.

The category of contrast/concession contains, according to the *Longman Grammar of Spoken and Written English*, 'items that in some way mark incompatibility between information in different discourse units, or that signal concessive relationships' (Biber et al. 1999:878). Interestingly, we found that the lexical bundle *at the same time* often signalled contrast, although its residual meaning suggests a temporal relationship (the phrase is given as an example of a temporal conjunction in Halliday and Hasan's summary table of conjunctive relations (1976:242–3)). Examples 6 and 7 illustrate the contrastive use of *at the same time*. In Example 6 the lecturer is discussing problems with Canada Geese. Contrast is signalled by *but*, and reinforced by the contrastive use of *at the same time.*

Example 6
They defecate in tremendous amounts, um, on sidewalks where runners are jogging around the lakes in the twin cities there's always a crisis about too many geese, but *at*

the same time, um, when they tried to do something about it, there were people who wanted them there, even in large numbers.

In Example 7 three parallel structures ('If … he says…'; 'If he says…'; 'if he then goes on to say…') introduce three examples of different modes of thinking. The bundle *at the same time* before the second example is used in the same way as *but* before the third example, to signal the contrast between simplex, complex and multiplex thinking.

Example 7
If you go to an historian an American historian and you say tell me about Abraham Lincoln and what happened in the theatre and he says the bullet the calibre of the bullet was so and so the assassin was called such and such this is what happened that's simplex. If he says at the same time there were twelve other people trying to assassinate him and security arrangements had been bunked up and the theatre had this he did okay that's pretty complex *but* if he then goes on to say but if Abraham Lincoln had not been Abraham Lincoln who had been elected but somebody else this would have happened if Abraham Lincoln had not been shot then this is what would I think would have happened to American history that's a bit of multiplex thinking.

Examples 8 to 11 show lexical bundles that signal result/inference. In Example 8 the two most frequent bundles in the corpus are combined: *the end of the* and *at the end of*. Typically these bundles did not serve as reference markers, being used in their literal sense to refer to points in time and place external to the lecture (*the end of the war, the end of the stage* etc.) or as a means of structuring and predicting events within the lecture itself (for example *I'll come back to it at the end of the lecture*). There were, however some examples of reference marking with the idiomatic expression *at/by the end of the day*. COBUILD defines this expression in the following way:

> You say at the end of the day when you are talking about what happens after a long series of events or what appears to be the case after you have considered the relevant facts. (*Collins COBUILD Advanced Learner's Dictionary* 2003)

Example 8 is interesting because the intention to communicate result/inference (what appears to be the case after you have considered the relevant facts) overrides the more residual meaning which might be paraphrased as 'finally' (what happens after a long series of events). The lecturer recommends that statistical advice should be taken '*at the end of the day*', as a logical consequence of the arguments put forward earlier, although the taking of statistical advice is also in fact the first step in the process of making a proposal.

Example 8
So you've got to be sure that when you make your proposal not only is your methodology right but you've got to have access to the cer-, to the to the facilities to do the work and at *the end of the day* a bit like statistical power really you ought to have statistical advice in other words take statistical advice before you start not at the end.

Example 9 describes a causal chain of events. The first causal link is not explicitly signalled: knowledge moves on and (implicit — *because of this*) you realise a mistake has been made. The second causal link is signalled with the cohesive conjunction *so*: you realise a mistake has been made (explicit — *so*) you go back and try to re-think. A third causal link is signalled by the bundle *to be able to*: you need the chain (explicit — *to be able to*) get back. The relationship between the latter two propositions could be restated in the following way — *You need the chain because you need to get back.*

Example 9
Knowledge moves on you realise that a mistake has been made way down there somewhere and so you go back and try to re-think you need the chain you need *to be able to* get back so what you leave in place is not that you are responsible for the whole pyramid but you are responsible for your brick and stone.

As in Examples 4 and 5, the lecturers in Examples 10 and 11 make use of two methods of signalling, both linking adverbials and lexical bundles. In Example 10 *hence* reinforces the function of the bundle *the end of the day*, and in Example 11 *so* reinforces the function of *in terms of.* In both cases the linking adverbials *therefore* or *consequently* could be substituted for the multi-word expressions.

Example 10
What we've done is we've included within their utility function the idea of everything that gives them satisfaction and *hence* at *the end of the* day they must allocate all of their incomes to those things that give them satisfaction including major saving

Example 11
You can commit serial murder at Warwick but plagiarism is beyond limits *so in terms of* the final piece of work it's your own OK

3.3.2 *Discourse organizers*

Some of the frequent bundles in our data appeared to be used to signal how the topic and/or activity of the discourse in one part of a lecture related to that in another. 'Transition', a category of linking adverbials in the *Longman Grammar of Spoken and Written English*, contains items signalling the introduction of discourse that is loosely connected or not connected to previous discourse. These adverbials "mark a transition to another, usually tangential, topic" (Biber et al. 1999:879). Signals of transition were relatively rare in all the registers investigated (conversation, academic prose, fiction and news reports); this is probably because the written registers could make use of typographical devices to signal topic change and asides, whilst in conversation the topics are not usually pre-planned, but are negotiated by all the participants.

Biber, Conrad and Cortes (2004:386) identified a number of 'topic introduction/ focus' discourse organizers which were common in their teaching subcorpus (occurring 40 to 99 times per million words) but were rare in other registers (occurring fewer that ten times per million words). These included *if you look at; take a look at; if you have a; if we look at; going to talk about; to look at the; to go ahead and; I want to do; what I want to; want to do is; want to talk about; you know if you; a little bit about*. The same or similar bundles also had this function in our lecture corpus. Example 12 shows the use of *if you look at* to signal transition. The end of discussion of the previous topic is

signalled by the linking adverb *finally*, and the start of discussion of the next topic is signalled by both the linking adverb *so* and the lexical bundle.

Example 12
and *finally* in this sort of hierarchy of grafting we have a xenograft and xeno means foreign and xenografts come from or or xenogeneic graft comes from a member of a different species for example pig to man yeah *so* let's *if you look at* the range of transplant medicine what is done and why i'm gonna talk briefly about currently successful grafts

As Biber, Conrad and Cortes (2004:382) point out, 'topic introducing bundles often result in syntactic blends'. In Example 12, *if you look at* is inserted mid sentence immediately following 'let's', and functions as if it were a non-finite verb form. As a result the sentence 'finishes up in a way that is syntactically inconsistent with the way it began' (Biber et al. 1999:1064).

Some of the other bundles Biber, Conrad and Cortes (2004) list as signalling topic introduction/focus were not common enough to be listed in the top 20 in our corpus, but occurred with a certain amount of frequency, for example *I want to do* (47 instances), *what I want to* (73 instances) and *want to do is* (37 instances). All of these seemed to function on occasion as signals of topic change. The more frequent cluster *I want you to* is categorised as a stance expression with a directive function by Biber, Conrad and Cortes (2004), rather than as a topic introduction/focus discourse organizer. In most of the examples in our corpus, however, it functioned simultaneously as a directive and a signal of activity change. It was frequently preceded by another marker such as *so* or *now*, as in Examples 13 and 14:

Example 13
So *I want you to* go through this diagram and make sure that you can do that for each step

Example 14
So now *I want you to* spend a few minutes, um, thinking about the consequences

Example 15 provides further evidence of the prefabricated nature of bundles. The speaker makes a false start, backtracks by inserting *I want you to*, and then resumes the directive.

Example 15
I'm going to ask you questions and think *I want you to* think back to France and Germany and to the Netherlands.

Example 16 illustrates the use of *I want you to* with the downtoner *a little bit.*

Example 16
okay operation mode which we'll look at in a minute and an effective address well *i want you to* do *a little bit* now i've been talking for twenty minutes about time you did something so write down all the possible forms all right for variants one and two okay.

In this example the lecturer is preparing the students to perform a task, and the use of *a little bit* may help to mitigate the threat to the audience's face by indicating that the imposition on their time and concentration will not be great.

The bundle *a little bit about*, listed by Biber, Conrad and Cortes (2004) as a common topic introduction/focus discourse organiser in classroom teaching, occurred 56 times in our lecture corpus but was not amongst the 20 most frequent bundle types. In almost every case it helped to signal topic change, as in Examples 17 and 18, or it functioned summatively to mark the end of a topic, as in Examples 19 and 20:

Example 17
Let's just think *a little bit about* this planar haem molecule

Example 18
I want to talk first before I stop *a little bit about* this issue of tissue matching

Example 19
I've talked today *a little bit about* the attributes of nationalism

Example 20
So that's *a little bit about* myself.

Biber, Conrad and Cortes (2004) list the variant *a little bit of* as a referential expression, linking propositions rather than topics or activities, but several of the instances in our corpus seemed to function to indicate topic change, often in conjunction with another signalling cluster. Examples 21 and 22 show the prospective use of *a little bit of*, proceeded by the frame markers *now* and *what I want to do now*:

Example 21
Now I'm going to dip into *a little bit of* my own research

Example 22
What I want to do now is to get back to *a little bit of* economics.

Example 23 illustrates retrospective use to signal the end of a digression, in conjunction with the frame markers *all right* and *so*:

Example 23
All right so that's *a little bit of* a sideway sideline.

Any sort of engineered change in topic or activity is essentially directive, and therefore potentially face threatening, and in all the above contexts the expression *a little bit* seems to help maintain face; either that of the audience, by downplaying the task imposition (Examples 17, 18, 21, 22), or that of the lecturer, by downplaying apparent digressions (Example 23, and possibly 20) and by giving the audience to understand that the information provided is only a small part of what there is to know and what the lecturer actually knows about the subject (Examples 18, 19 and 22). Similarly in Examples 20 and 21 the lecturers mitigate a potential threat to their audience's negative face by using *a little bit* to downplay talk about themselves and their own contribution.

The bundle *in terms of the* is also listed by Biber, Conrad and Cortes (2004) as a referential expression, like *a little bit of.* In our corpus *in terms of the* sometimes seemed to be used to establish logical relationships in the text (as in Example 11) but also seemed to be used to mark a transition between topics, as in Example 24. In this example, following an aside by the lecturer on his or her own ways of coping with study, work and family as a student, *in terms of the* is used to signal a return to the main topic, the assignment. The bundle appears to signal retrospectively that the preceding discourse was tangential to the main discourse.

Example 24
I survived an MBA a family and a job purely by working with a group of mates here we sorted out the competencies of each of us the time of each of us and sort of divvied up the work. *In terms of the* presentation of the final paper, that's an individual piece of work and must be handled as such.

Finally, the following examples show the use of *I'm not going to* prospectively, to narrow the focus of discussion (Examples 25 and 26), and summatively, to signal the end of a lecture topic (Examples 27 and 28).

Example 25
Yeah now this is what this says and *I'm not going to* talk about the second equation but just the first one

Example 26
I told you last time because I didn't finish the lecture on Rawls and utilitarianism that I was going to finish it today but actually I've decided *I'm not going to* I know you'll be very disappointed but I think you've had enough of that so I'm going straight into Nozick and his criticisms of Rawls

Example 27
You can get some idea of the w- the snow water content there and *I'm not going to* do any more and talk about this

Example 28
I've set out my defence there — you've got it *I'm not going to* say anything more.

4. Summary and conclusion

The analysis in the previous section indicates that four word lexical bundles can play a discourse signalling role in lectures, and we would argue that it is important for language learners to be aware of this. While native speakers of English can be expected to have implicit knowledge of the function of bundles, non native speakers are much less likely to have this understanding because they have consciously learned the language, rather than acquired it, and the role of lexical bundles as discourse signals is yet to be acknowledged in most language teaching materials.

In lectures students are required to process relatively long stretches of discourse featuring a complexity of ideas, discourse topics and activities. Discourse signals are intended to help the listener predict the nature of upcoming ideas and information, and

a student who is unable to recognize these signals will be faced with additional cognitive processing demands, having to deduce both the intrinsic meaning of propositions, and make inferences about the relations between them.

Experimental studies have indicated that discourse signals in written text affect a reader's ability to comprehend the text and recall information from it (Martinez 2002). Hoey (2001) argues that readers formulate hypotheses about how written text will develop, and this helps them interpret it correctly. According to Hoey (2001:32) "accurate recognition of the signals and their significance for the text's development can greatly ease a reader's processing burden in that it lessens the need for large-scale hypothesis forming at the same time as all that detailed micro-processing going on". We would argue that this is also the case when listeners interact with lectures.

A considerable body of research has focused on the devices used in written text to make the links between sentences unambiguous for the reader (a recent example being Cortes 2004). The devices used in spoken text have been less extensively researched, however, and only limited information about them is available to teachers. Parrot (2000) makes some attempt to offer such a description, including, for example, items in 'general use' for signalling exemplification (*for example*, *for instance* and *e.g.*) and providing one item 'used mainly in speaking' (*say*). It is clear that more data-driven descriptions of the features of spoken language are needed to supplement the lists of discourse markers typically presented to learners of English as a foreign language.

Some writers (Flowerdew 2003; Thurstun & Candlin 1997, 1998; Weber 2001) suggest an inductive approach to teaching vocabulary use, in which students employ corpus-based research methods, referring to on-screen concordances (or print outs from them) to examine how words are used in context. This might also be a suitable way to teach students about discourse organising devices. If applied to spoken discourse, such an approach would entail providing transcriptions for students to examine, or requiring the students themselves to transcribe pre-selected excerpts from recordings (such as the examples given in Section 3 of this paper). They might then be required to identify the ideas and topics in the excerpts, and discuss the relationships between them and the means whereby these relationships are signalled.

As far as we are aware, the cohesive role of lexical bundles has not yet been examined in the classroom in this way. Now that lists of frequently occurring bundles are becoming available, however, we think that it is useful for learners to investigate their use as signalling devices in authentic text, alongside the better known exponents that have come to feature prominently in English language teaching materials since the publication of the seminal work of Halliday and Hasan (1976).

1. The BASE corpus is a collection of academic speech events under development at the Universities of Warwick and Reading with funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research
Board http://www.warwick.ac.uk/go/base/

2. MICASE is on-line, searchable collection of transcripts of academic speech events recorded
at the University of Michigan. http://www.lsa.umich.edu/eli/micase/index.htm

References

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word combinations. In A. P. Cowie (Ed.), *Phraseology* (pp. 101–122). Oxford: Oxford University Press.

Barlow, M. (2002). *MP 2.2*. Houston, TX: Athelstan Software.

Banerjee, S. & Pedersen, T. (2003). The Design, Implementation, and Use of the Ngram Statistics Package. *Lecture Notes in Computer Science*, Vol 2588, 370–381.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E. & Urzua,A. (2004). *Representing language use in the University: Analysis of the TOEFL 2000 Spoken and Wrtitten Academic Language Corpus*. TOEFL Monograph Series MS 25.Princeton, New Jersey: Educational Testing Service.

Biber, D., Conrad, S. & Cortes, V. (2004). *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25 (3), 371–405.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman

Collins COBUILD Advanced Learner's English Dictionary. (2003). Sinclair, J. (Ed.). Glasgow: HarperCollins Publishers.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes*, 23 (4), 397–423.

DeCarrico, J. & Nattinger, J. (1988). Lexical Phrases for the comprehension of academic lectures. *English for Specific Purposes*, 7, 91–102

Flowerdew, J. (2003). Signalling nouns in discourse. *English for Specific Purposes*, 22, 329–346.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English* London: Longman.

Hoey, M. (2001). *Textual Interaction*. London: Routledge.

Hyland, K. (2004). Disciplinary interactions: metadiscourse in L2 postgraduate writing.*Journal of Second Language Writing*, 13 (2), 133–151.

Louwerse, M. M. & Mitchell, H. H. (2003). Towards a taxonomy of discourse markers in dialogue: A theoretical and computational linguistic account. *Discourse Processes*, 35 (3), 199–239.

Martinez, A. C. L. (2002). Empirical examination of EFL readers' use of rhetorical information.*English for Specific Purposes, 21*, 81–98.

Nattinger, J. & J. DeCarrico (1992). *Lexical Phrases and Language Teaching.* Oxford: Oxford University Press.

Parrot, M. (2000). *Grammar for English Language Teachers*. Cambridge: Cambridge University Press.

Scott, M. (1997). *WordSmith Tools Manual.* Oxford: Oxford University Press

Sinclair, J. & Coulthard, R. M. (1975). *Towards an Analysis of Discourse: the English used by teachers and pupils.* Oxford: Oxford University Press

Strzalkowski, T. (Ed.) (1998). *Natural Language Information Retrieval* (Introduction pp. ixx).Dordrecht: Kluwer

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, 7 (2), 215–244.

Thurstun, J. & Candlin, C. N. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes Journal, 17* (3), 267–280.

Thurstun, J. & Candlin, C. N. (1997). *Exploring Academic English: A Workbook for Student Essay Writing*. Sydney: National Centre for English Language Teaching and Research, Maquairie Univesity.

Weber, J. J. (2001). A concordance and genre-informed approach to ESP essay writing. *English Language Teaching Journal*, *55* (1), 14–20.